

Combinatorial Experimental Strategies

Z.Q.

John Lu

May 22, 2003

NCMC Combi Workshop

Statistical Engineering Division

National Institute of Standards and Technology

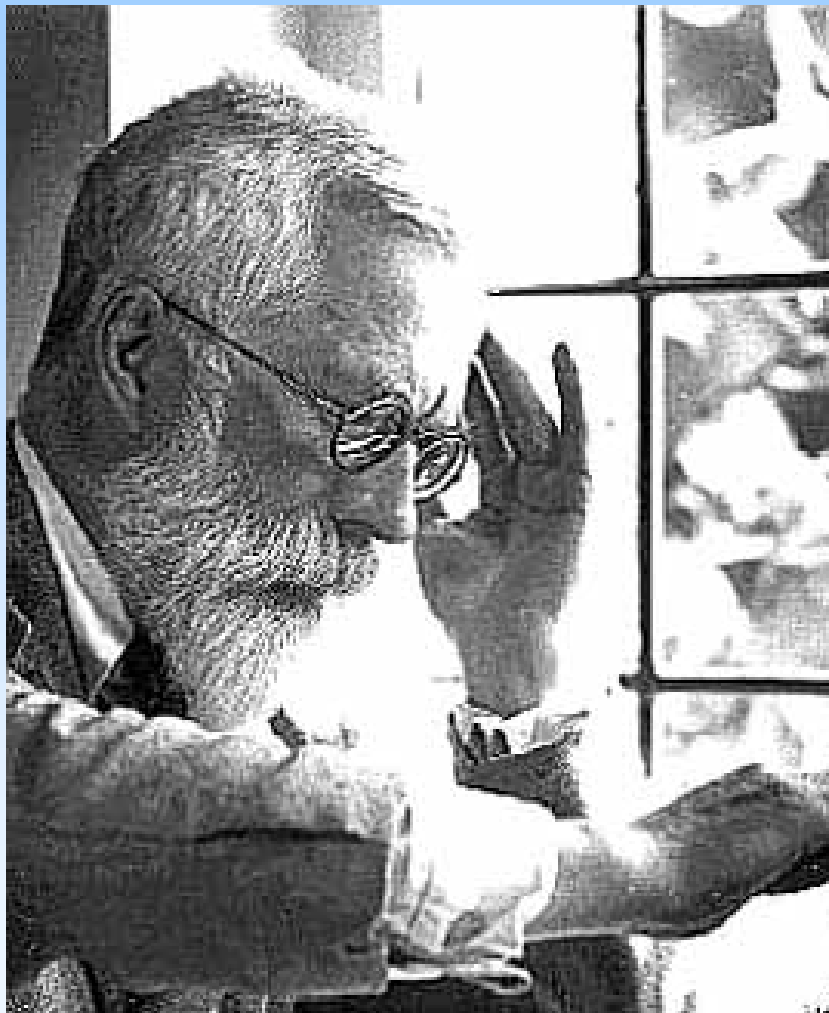
Why Do We Need Statistical Design?

- Design directly impact how and what information is collected (Design and choice of library at each experimental run! Meta info for informatics database.)
- The issue of exploratory vs exploitation (optimizing and searching) (How to design sequential experiment is crucial for Combi experiments.)
- Informatics for Combi is challenging both in terms of the amount of data generated, but also the speed of processing and organizing data. (Compress and collect only the most needed info!)

Why statistical approach in design?

1. Systematic approach: complete enumeration of all factors not feasible! Thus, need for combinatorial design: testing on subsets chosen randomly out of the whole.
2. Reduce effects from nuisance factors: Use of randomization removes systematic bias!
3. There is need for replication! Understand the reliability of data and repeatability!
4. Statistical approach allows quantifying confidence (uncertainty) in conclusion----Setting standards.

Part One: Combinatorial Design: Building Discrete Gradient Library



- R A Fisher (1890-1962): biologist (genetics) and father of mathematical statistics.
- Agri origin: Rothamsted Station, UK (starting 1919, 1926 paper, Design of Experiments Book 1935)
- Contributions:
 - Randomization
 - Replication
 - Blocking to reduce noise and power (false negatives)
 - Balance/orthogonality
 - Analysis procedure of variability: ANOVA

R A Fisher, as a scientist and statistical founder

“Perhaps the most original mathematical scientist of the [twentieth] century” **Bradley Efron** *Annals of Statistics* (1976)

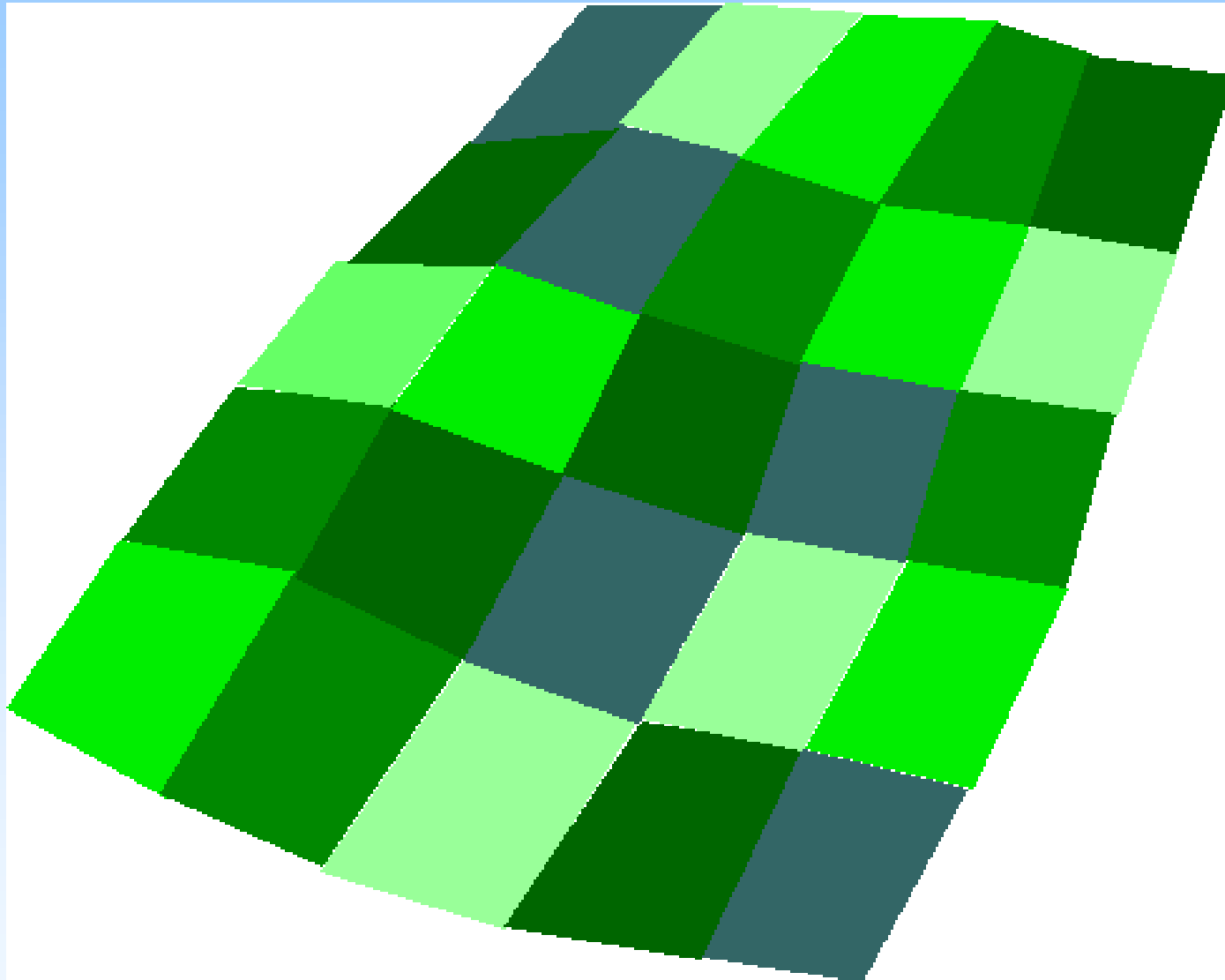
“Fisher was a genius who almost single-handedly created the foundations for modern statistical science” **Anders Hald** *A History of Mathematical Statistics* (1998)

“Sir Ronald Fisher ... could be regarded as **Darwin’s greatest twentieth-century successor.**” **Richard Dawkins**, *River out of Eden* (1995)

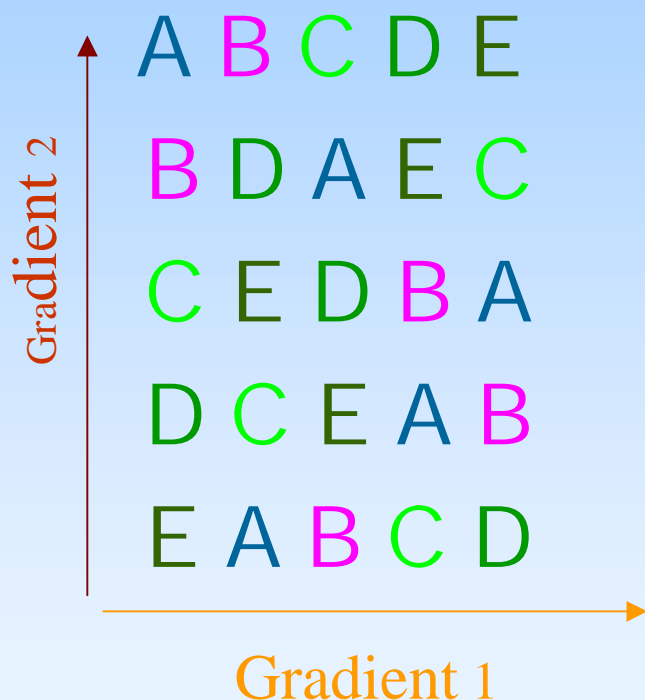
“When I was a student, there were a few contemporary scientists whose work and abilities I particularly admired: John von Neumann, R.A.Fisher, and Robert Oppenheimer.” **John H Holland** 1995, *Hidden Order: How Adaptation Builds Complexity*.

Joint Statistical Meetings 2003, Fisher Lecture: “**On rereading L.J. Savage rereading R. A. Fisher**”, by **A.F.M. Smith, FRS**

What is a Latin Square?



Latin Squares




- Idea:** Divide a large plot (land) into 5 x 5 grid of subplots (subareas, or **arrays**) and apply fertilizers A, B, C, D, E. (called treatments, can be catalysts, additives, that are of interest.)

- Randomization:** random labeling of fertilizers

- Balance:** Every label appears exactly once in each row and each column.

Comparison to systematic and randomized block design



| | | | | |
|---|---|---|---|---|
| A | B | C | D | E |
| A | B | C | D | E |
| A | B | C | D | E |
| A | B | C | D | E |
| A | B | C | D | E |

Systematic Design

| | | | | |
|---|---|---|---|---|
| A | C | D | E | B |
| B | D | C | A | E |
| C | E | B | A | D |
| D | E | B | C | A |
| D | C | B | E | A |

Randomized block design

Data decomposition for Latin square design

ANOVA = Arrangement of Data

| | | Cars | | | | | | | |
|---------|-----|------|------|------|------|---|----------|---|----------------|
| | | 1 | 2 | 3 | 4 | | | | |
| Drivers | I | A 21 | B 26 | D 20 | C 25 | = | Baseline | + | Rows (drivers) |
| | II | D 23 | C 26 | A 20 | B 27 | | | | |
| | III | B 15 | D 13 | C 16 | A 16 | | | | |
| | IV | C 17 | A 15 | B 20 | D 20 | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |

Comments on non-replicated design

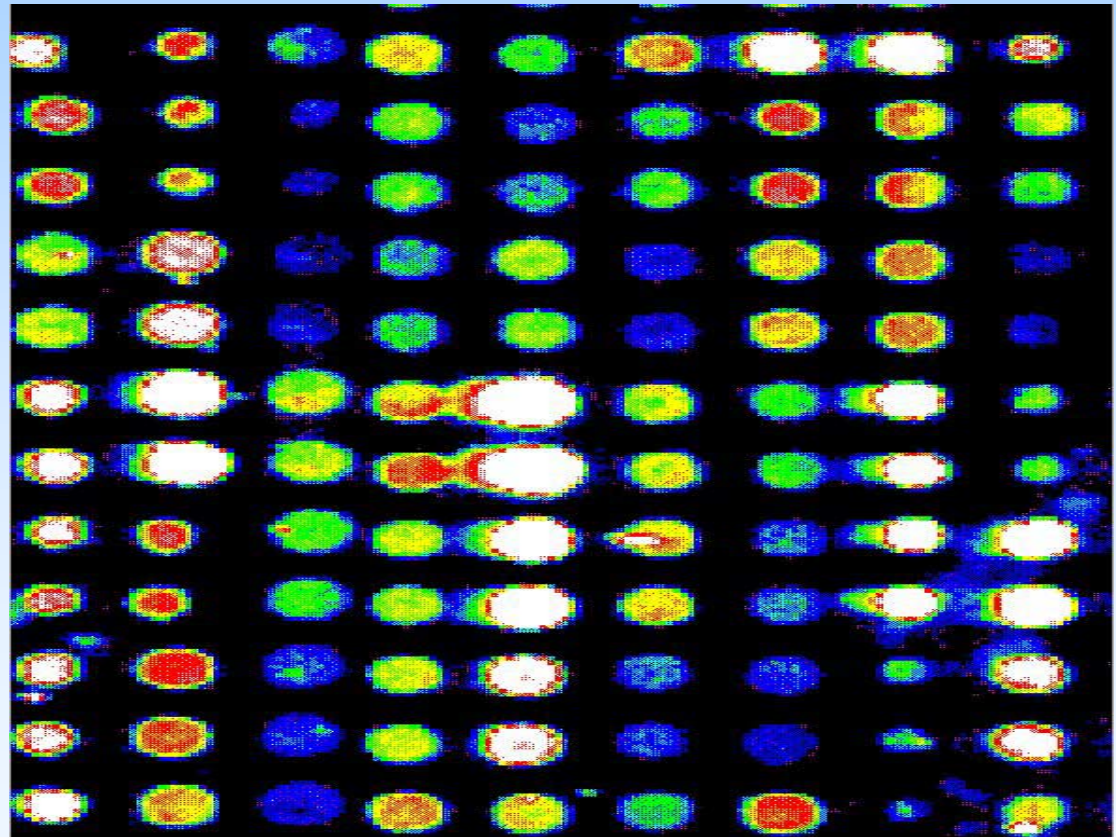
- Difficult to interpret data without strong model assumptions, i.e. the two-way block effects do not interact and the effect of treatment is **additive**.
- **Nonlinear normalization**: instead of row or column based normalization, more complicated transformation may be needed to remove position effects.

Excursion: Microarray Data Analysis

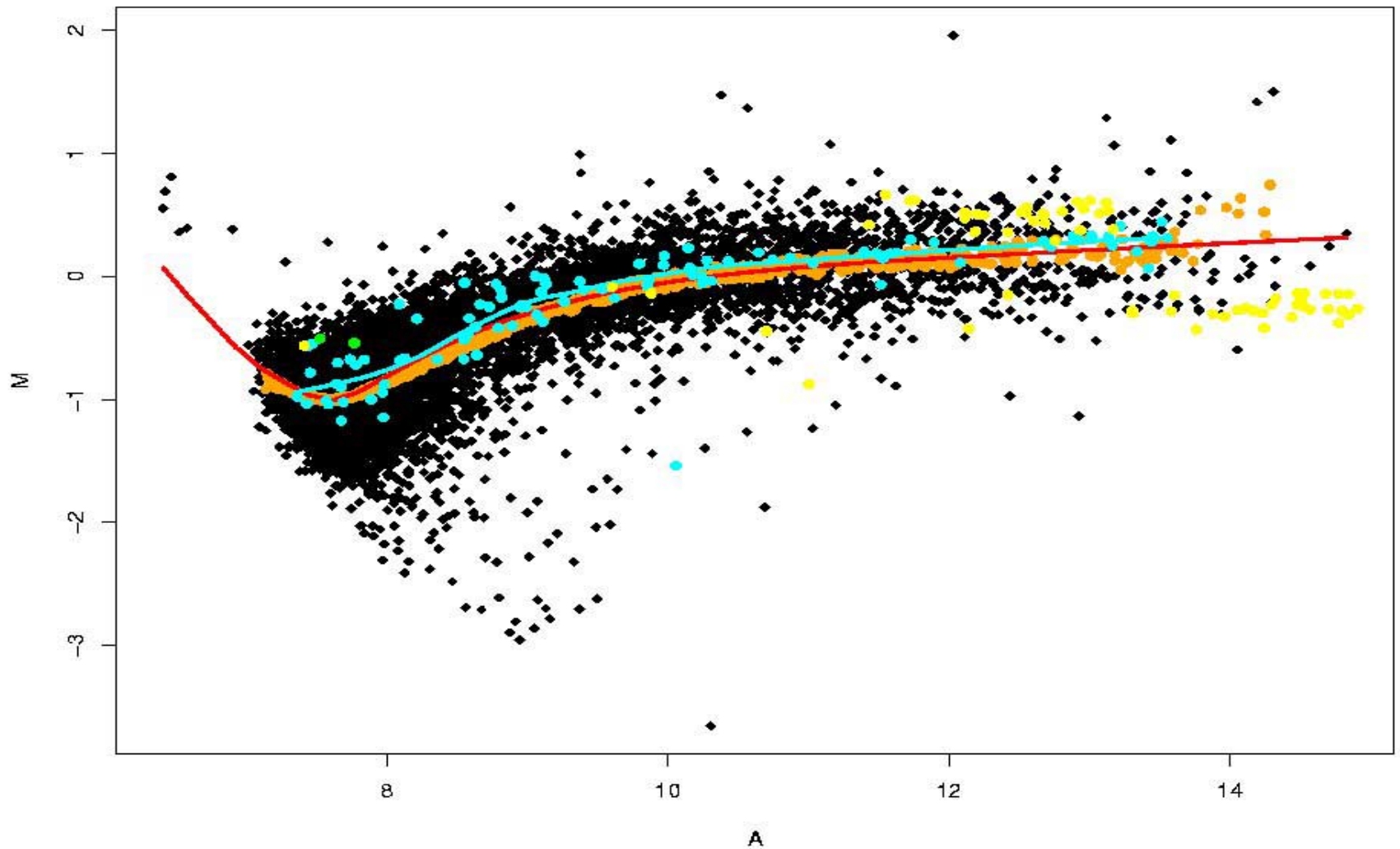
DNA microarray experiments: allow monitoring expression levels of thousands of genes simultaneously

Notations: for a spot I , $j=1,\dots,p$, let R_j and G_j denote the measured fluorescence intensities for red and green dyes.

Normalization: Let $M=\log_2(R/G)$, and $A=(1/2)*\log_2(RG)$,
Nonlinear Model $M=f^{\wedge}(A)$,
Compute $M^{\wedge}=M - F^{\wedge}(A)$,
which is the estimated gene effect.



Excursion: Nonlinear normalization



Orange: Schadt-Wong rank invariant set

Red line: lowess smooth

Excursion: Issues in microarray data analysis

- **Image analysis:** addressing, segmenting, quantifying
- **Normalisation:** within and between slides
- **Quality:** of images, of spots, of (log) ratios
- **Multiple comparison:** Which genes are (relatively) up/down regulated? (What are the hits?)
- Assigning p-values to tests/**confidence** to results.
- **Planning of experiments:** design, sample size
Etc.

Excursion: Standards in Microarray Experiments

- Minimum information about a microarray experiment-MIAME
 - www.mged.org
 - Check list for publications concerning microarray experiments
- Normalization, data quality, experimental design, and statistical analysis.



Back to General Design Principles

The general rule is:

"Block what you can, randomize what you cannot."

---- **Blocking** is used to remove the effects of a few of the most important nuisance variables.

----**Randomization** is then used to reduce the contaminating effects of the remaining nuisance variables.

-----**Replication** to get uncertainty estimate and of interaction.

Part two: Combi Experiments

Characteristics:

- **Interaction** is important
- Potentially many factors: 3-7
- **Parallelism**: generating a large number of diverse molecules in a template and studying them simultaneously in a library
- Response surface is complicated, high-dimensional, and unknown

Example: Fractal masking strategy to design combinatorial library

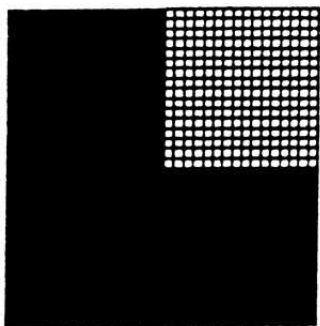
J. Wang, Y. Yoo, C. Gao, I. Takeuchi, X. Sun, H. Chang, X.D. Xiang, P.G. Schultz (1998), *Science*.

The sequence of masking and precursor deposition:

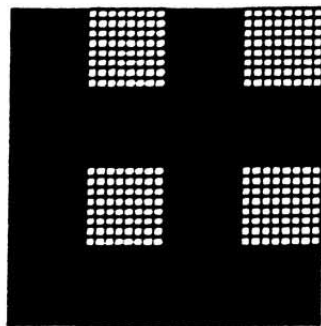
A_1 : Ga_2O_3 (355 nm); A_2 : Ga_2O_3 (426 nm); A_3 : SiO_2 (200 nm); A_4 : SiO_2 (400 nm);
 D_1 : CeO_2 (3.5 nm); D_2 : EuF_3 (11.3 nm); D_3 : Tb_4O_7 (9.2 nm); E_1 : Ag (3.8 nm); E_2 : TiO_2 (6.9 nm); E_3 : Mn_3O_4 (5.8 nm); B_1 : Gd_2O_3 (577 nm); B_2 : ZnO (105 nm); B_3 : ZnO (210 nm); C_1 : Gd_2O_3 (359 nm); C_2 : Y_2O_3 (330 nm); and C_3 : Y_2O_3 (82.5 nm).

The libraries consist of individual sites $650\text{ }\mu\text{m}$ by $650\text{ }\mu\text{m}$, spaced $100\text{ }\mu\text{m}$ apart, deposited on thermally oxidized Si substrates 2.5 cm square.

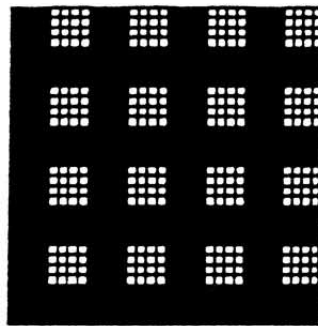
Masks used to generate the quaternary library



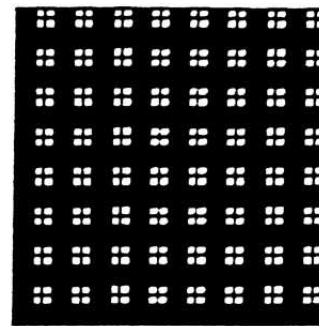
A_i



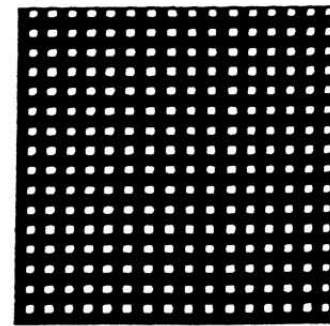
B_i



C_i



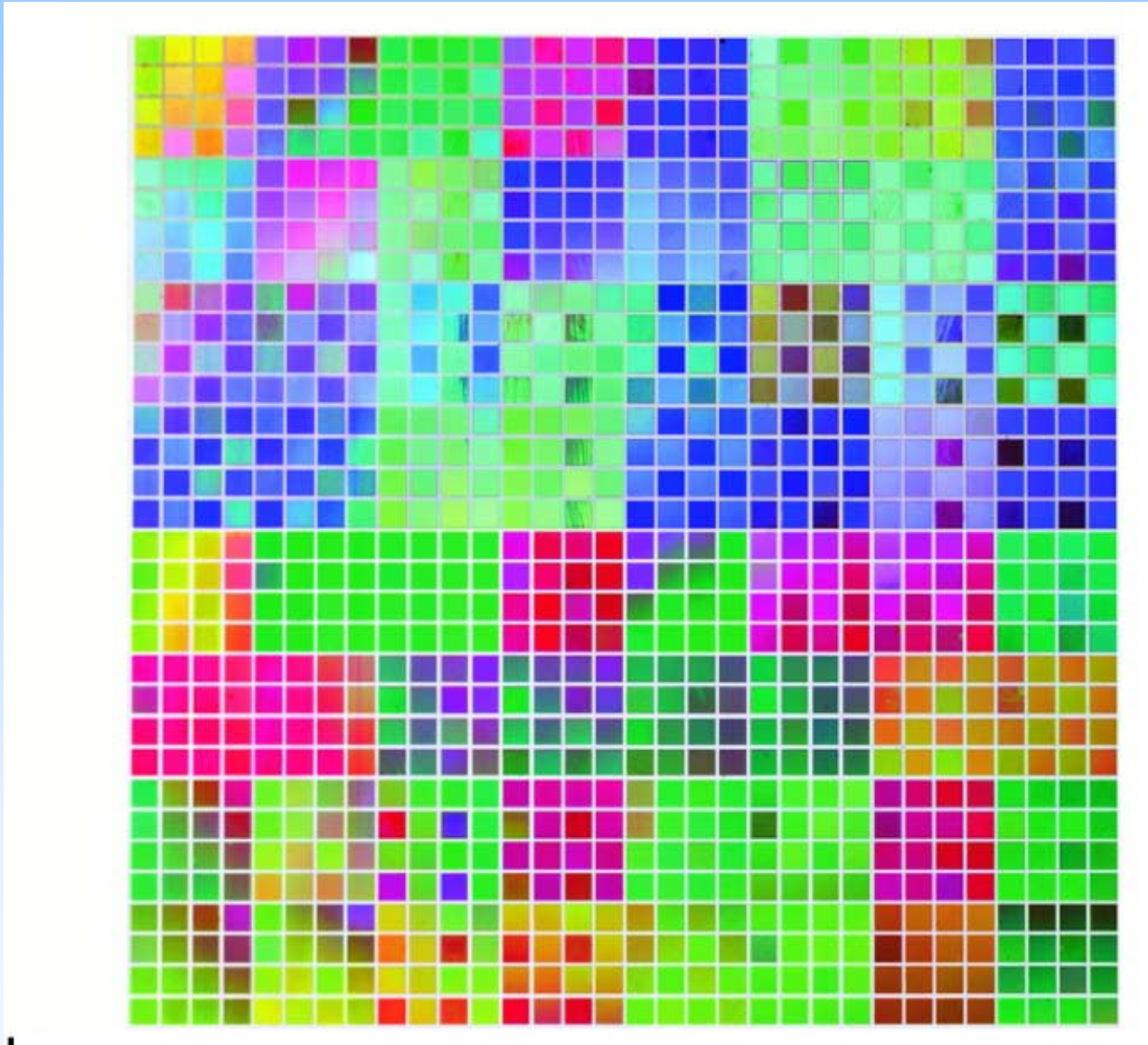
D_i



E_i

Masks for generating the quaternary library. A_i , B_i , C_i , D_i , and E_i represent a deposition step with mask X rotated counterclockwise by $(i - 1) \times 90^\circ$.

Quaternary library (32x32 square)



Comment:

- It is a Latin square shape, but the structure is much more complicated.
- Nested structure design
- Statistical analysis??

Other Useful Designs

- Balanced Incomplete Design (BIBD), esp. Youden square.
- Design for modeling building and optimization: response surface design and sequential design (G E P Box, Hunter, Hunter 1978)
- Optimal design: J. Kiefer (1959)
- Bayesian design criterion (D V Lindley): most suitable for sequential implementation and complicated modeling problems.
- Robust Design for noisy process control (Taguchi 1986-87, Wu and Hamada 2000)

Part three: Challenges and Proposals

- Can high throughput allows us to study higher-order interaction? Can we modify classical combinatorial design to deal with higher-order interaction
- Classifcal designs are low-throughput (small number of runs), we need to take them to high throughput (many runs) (R C Bose may be smiling!)
- The issues of array-to-array consistency, blocking, nested design structure, etc.

High-dimensional nonlinear response surface

- Example: quantitative structure property analysis
- Curse of dimensionality: multi-component and higher-order interaction (mixtures)
- Complicated response: stochastic process surface (?)
- Intrinsic low-dimensional structure (descriptors): dimension reduction such as SVD and PCA will be helpful!
- Singular design model: low-dimensional structure is a good thing even for nonparametric nonlinear modeling. Lu (1999), J. Multivariate Analysis
- SVD-based multivariate locally weighted regression (Lu 2003)

Late-stage design with chemical knowledge

- Exploratory design: Spatial design

Lu, Berliner, Snyder 2000, Springer Book

- Optimization /adaptive design: Bayesian sequential design with data assimilation process (Berliner, Lu, Snyder 1999, J. Atmospheric Sciences.)

Important references

1. Joan Fisher Box (1978): *R A Fisher: The Life of a Scientist*
2. G.E.P. Box, W.G. Hunter, J.S. Hunter (1978). *Statistics For Experimenters: An Introduction to Design, Data Analysis, and Model Building.*
3. D. Raghavarao (1971): *Constructions and Combinatorial Problems in Design of Experiments.*
4. J. N. Cawse (eds 2003): *Experimental Design for Combinatorial and High Throughput Materials Development.*